Recent Trends in Operating Systems and their Applicability to HPC

Arthur Maccabe,
Patrick Bridges

Currently:
Director, Computer Science
and Mathematics Division
Oak Ridge National Laboratory

Ron Brightwell,
Rolf Riesen

University of New Mexico

The University of New Mexico

Sandia National Laboratories

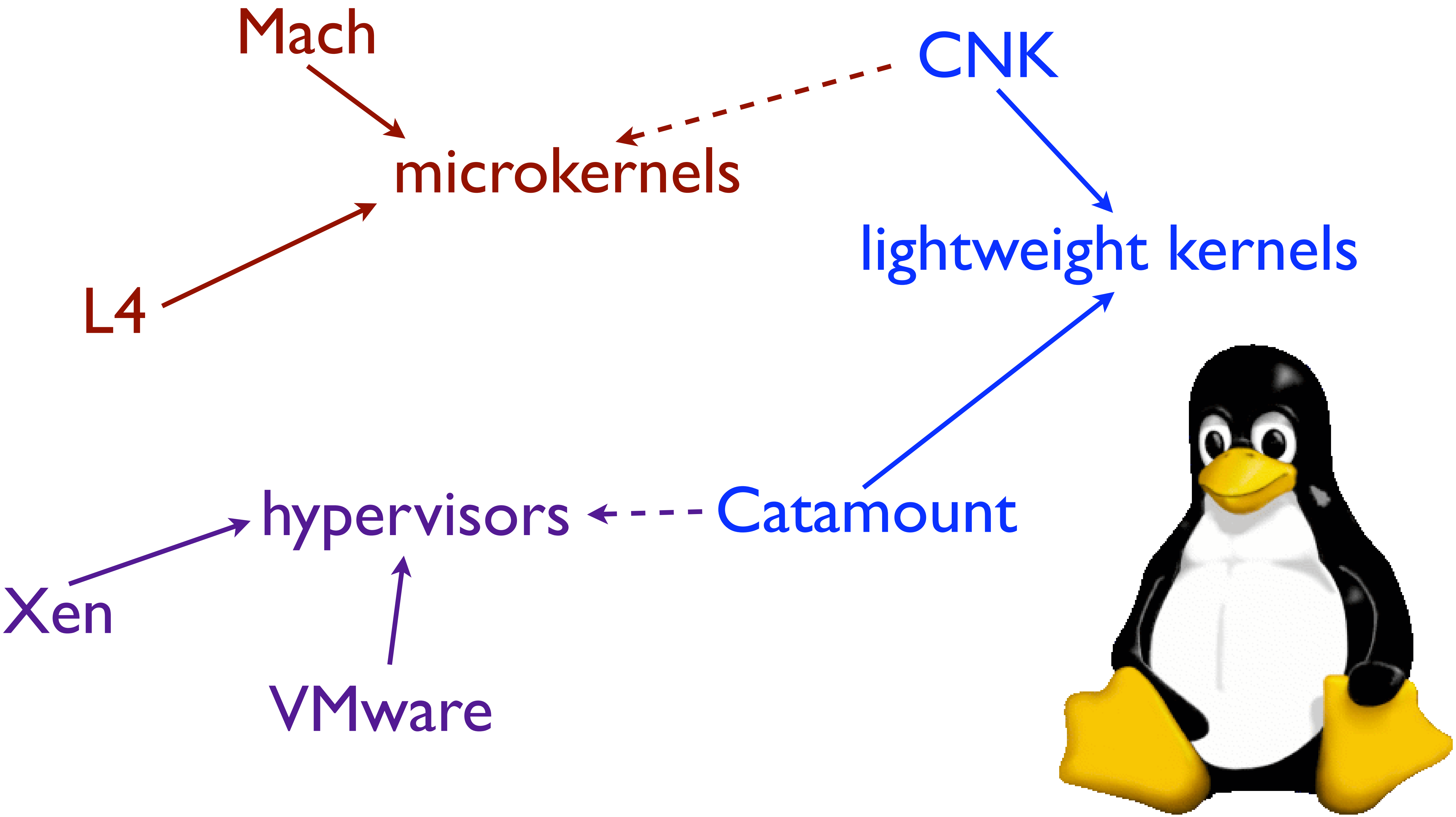
Sandia
National
Laboratories

Scaling to New Heights
CUG 2006

May 11, 2006
Lugano, Switzerland

CRAY USER GROUP
INCORPORATED

Variety, variety, variety



Mach

CNK

microkernels

lightweight kernels

L4

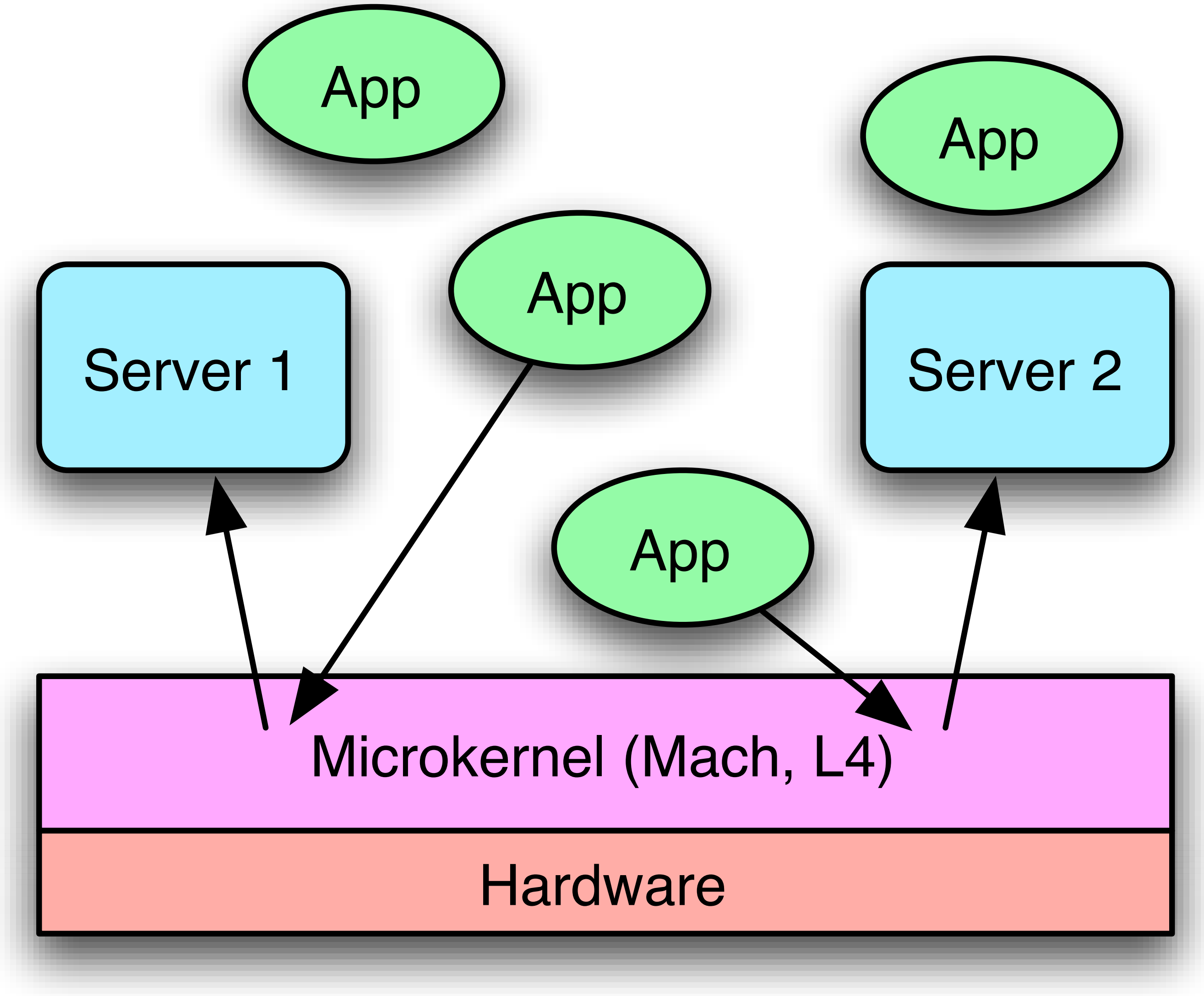
hypervisors

Catamount

Xen

VMware

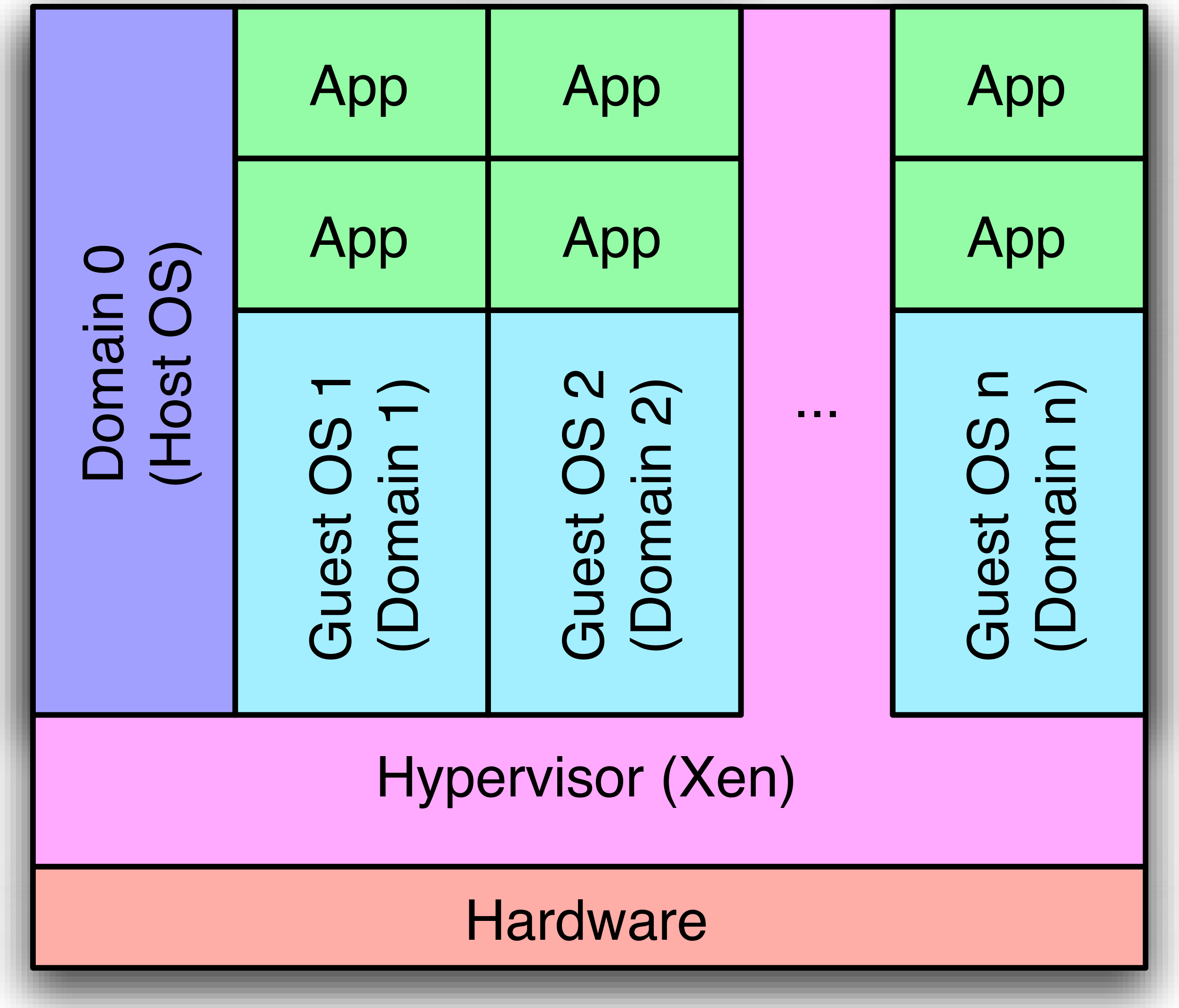
Microkernels



- Minimal services
  - policy versus mechanism
  - address spaces, control (threads), message passing
- Servers
  - trampoline

Hypervisors

- Hypervisor virtualizes hardware
  - goal is to run multiple OSes
  - direct access to hardware is preferred
- Xen (para)virtualizes Processor, MMU, and basic I/O
  - Additional I/O virtualization done by Domain 0



Domain 0 (Host OS)

App App App

App App App

Guest OS 1 (Domain 1)

Guest OS 2 (Domain 2)

...

Guest OS n (Domain n)

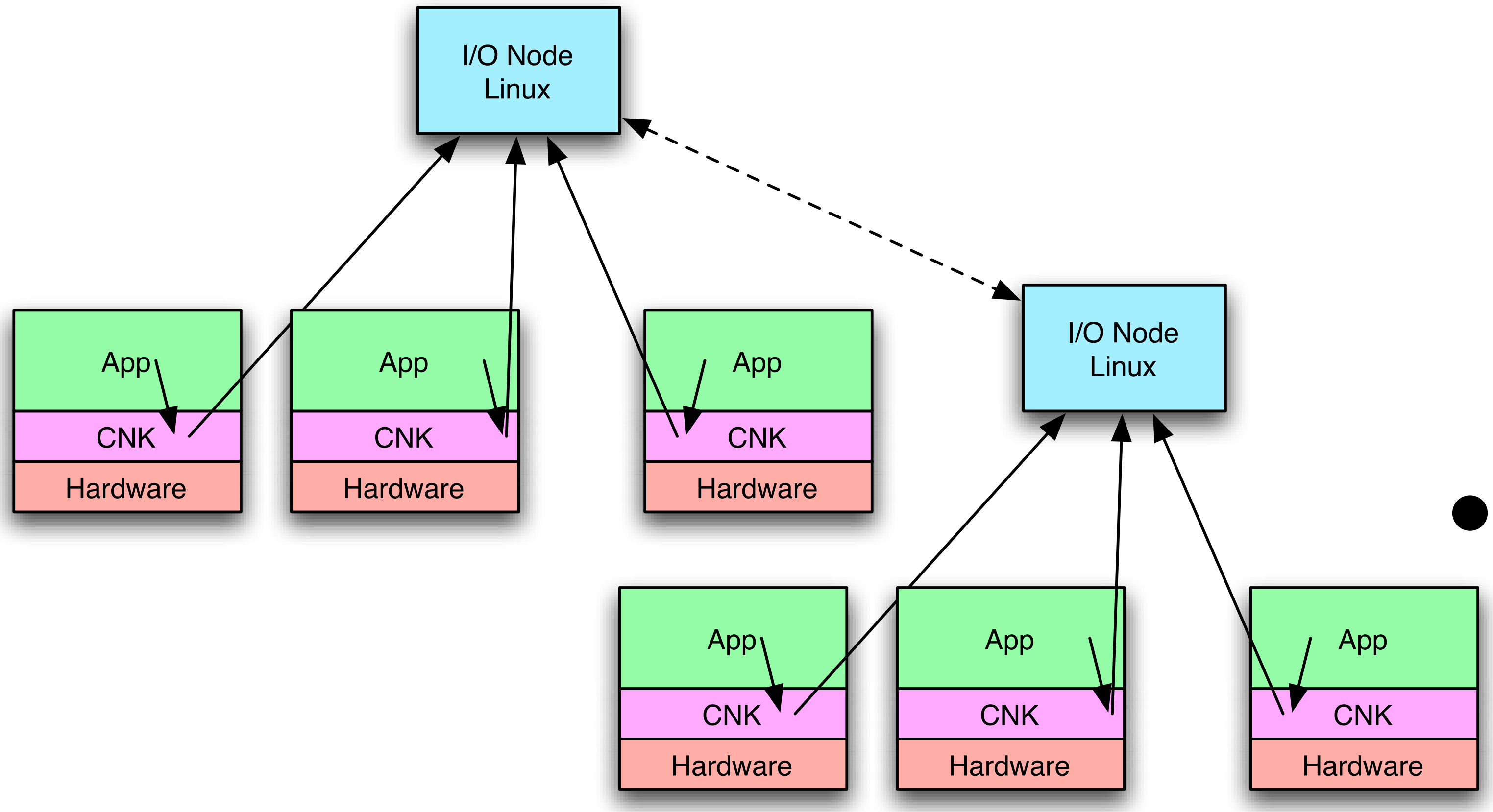
Hypervisor (Xen)

Hardware

Lightweight Operating Systems

- Catamount
  - SUNMOS, Puma/Cougar
  - Catamount, Portals
- Blue Gene/L
  - Compute Node Kernel (CNK)
  - I/O Nodes (linux)

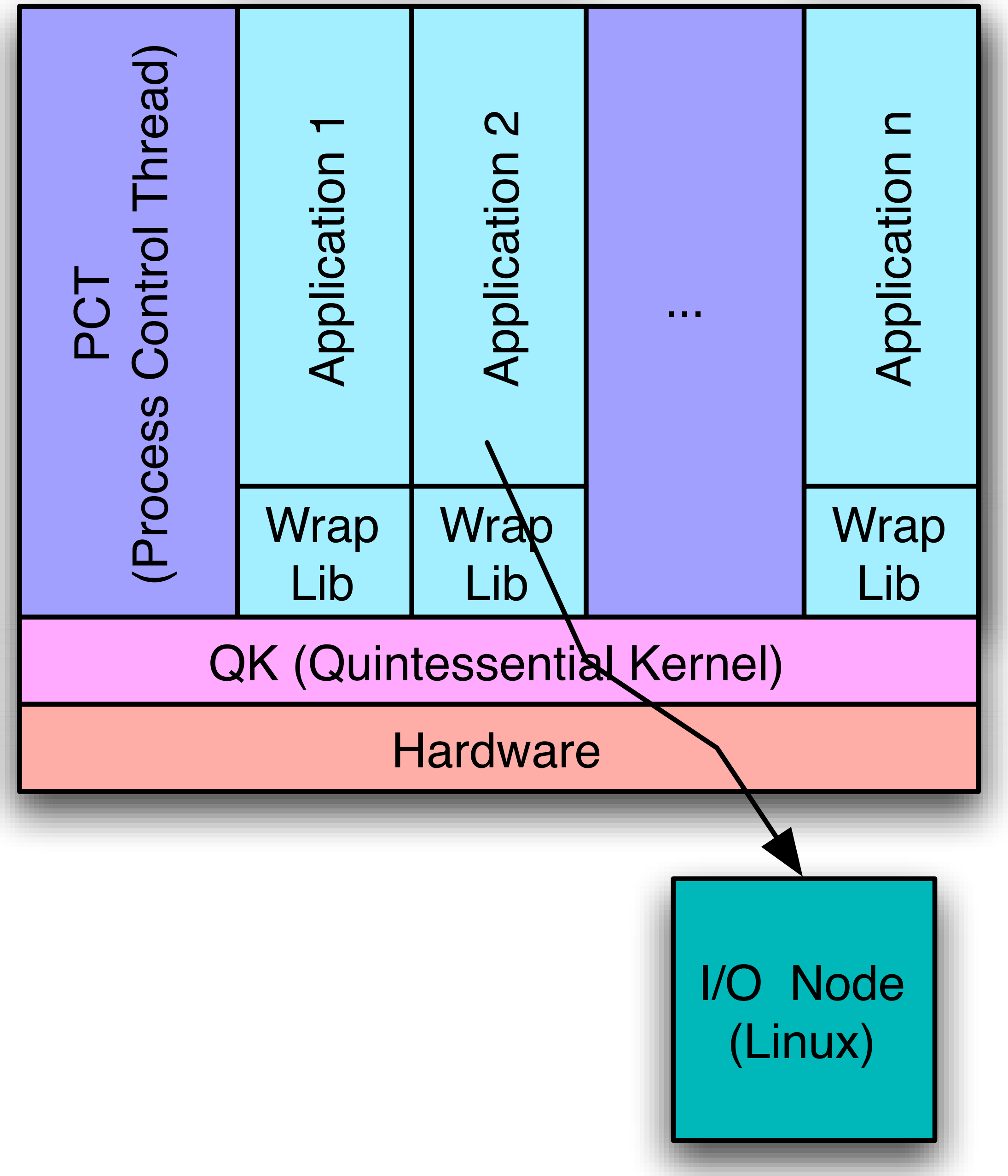
Blue Gene/L CNK



- I/O nodes
  - run Linux
  - have storage resources
  - separate I/O network
- Compute nodes
  - run lightweight kernel
  - high speed, partitionable network

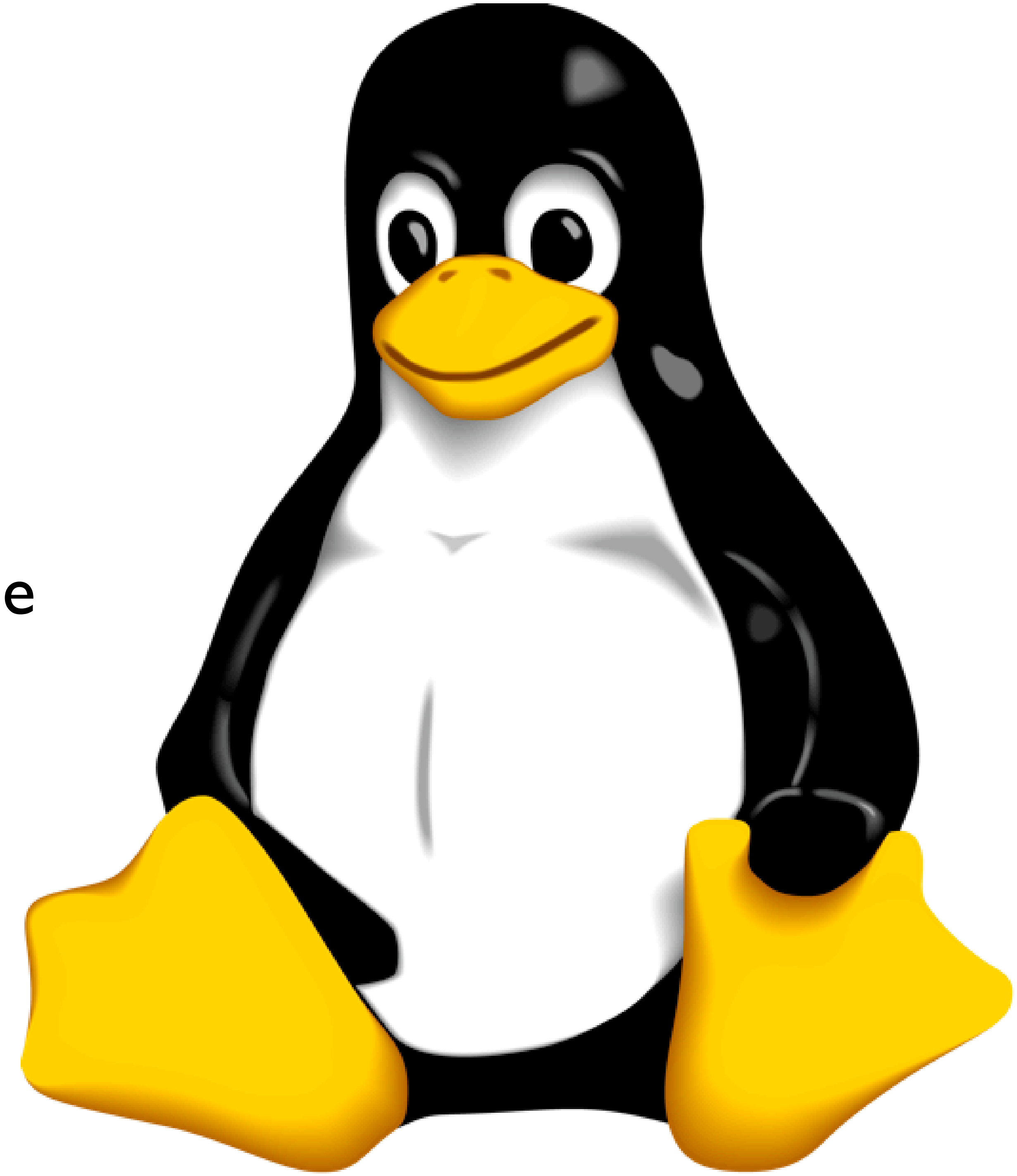
Catamount

- QK – mechanism
  - communication
  - address spaces
- PCT – policy
  - finding servers
- Wrapper lib
  - wrapper for stdio calls
  - RPC to I/O node



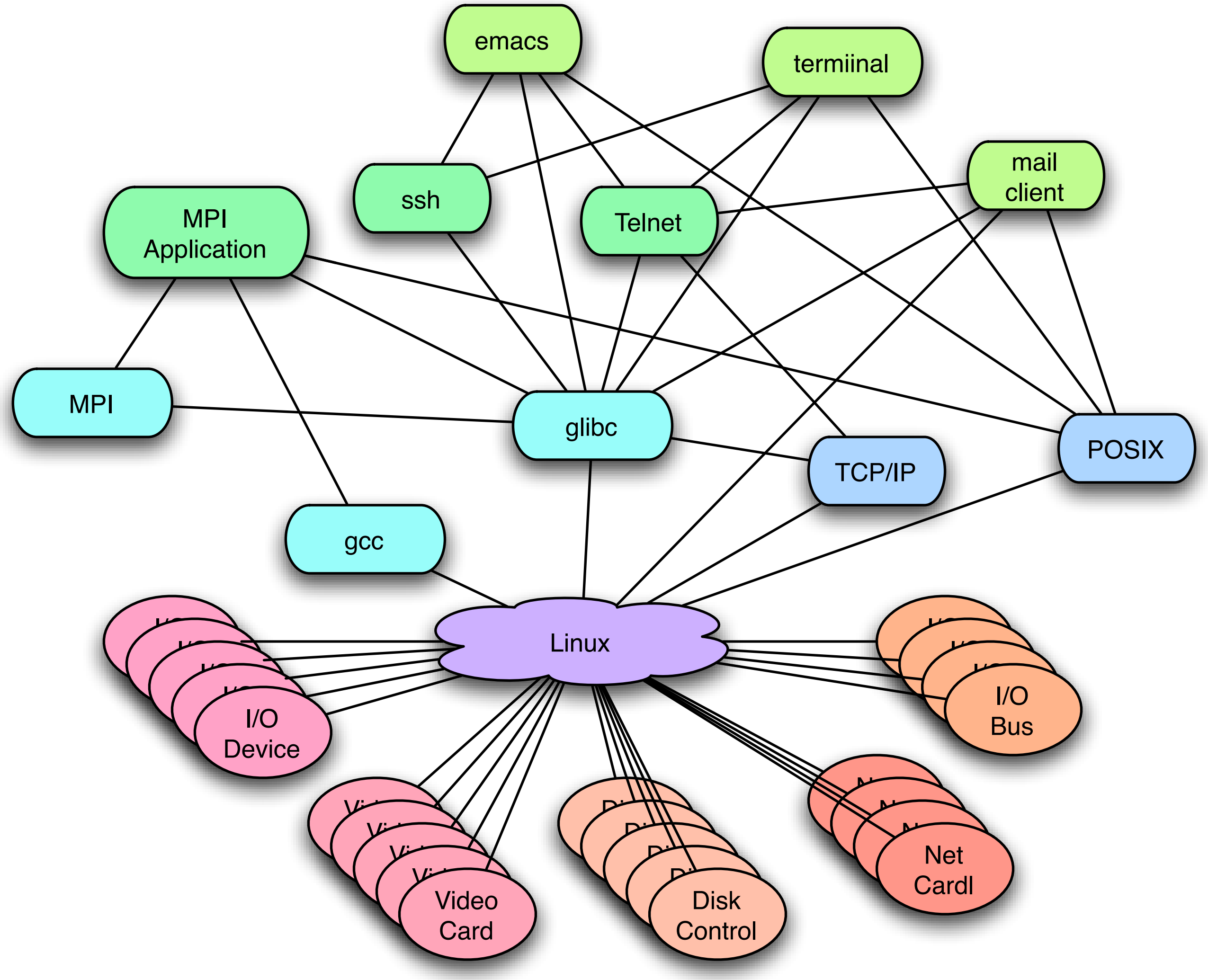
Linux, the 800 Pound Penguin

- Imagine that you are a “small” computer company in the US
- One customer believes in lightweight OSes
- Another demands Linux
- You can’t afford to support the code bases for two OSes
- What do you do?
- The world is waiting for your answer....

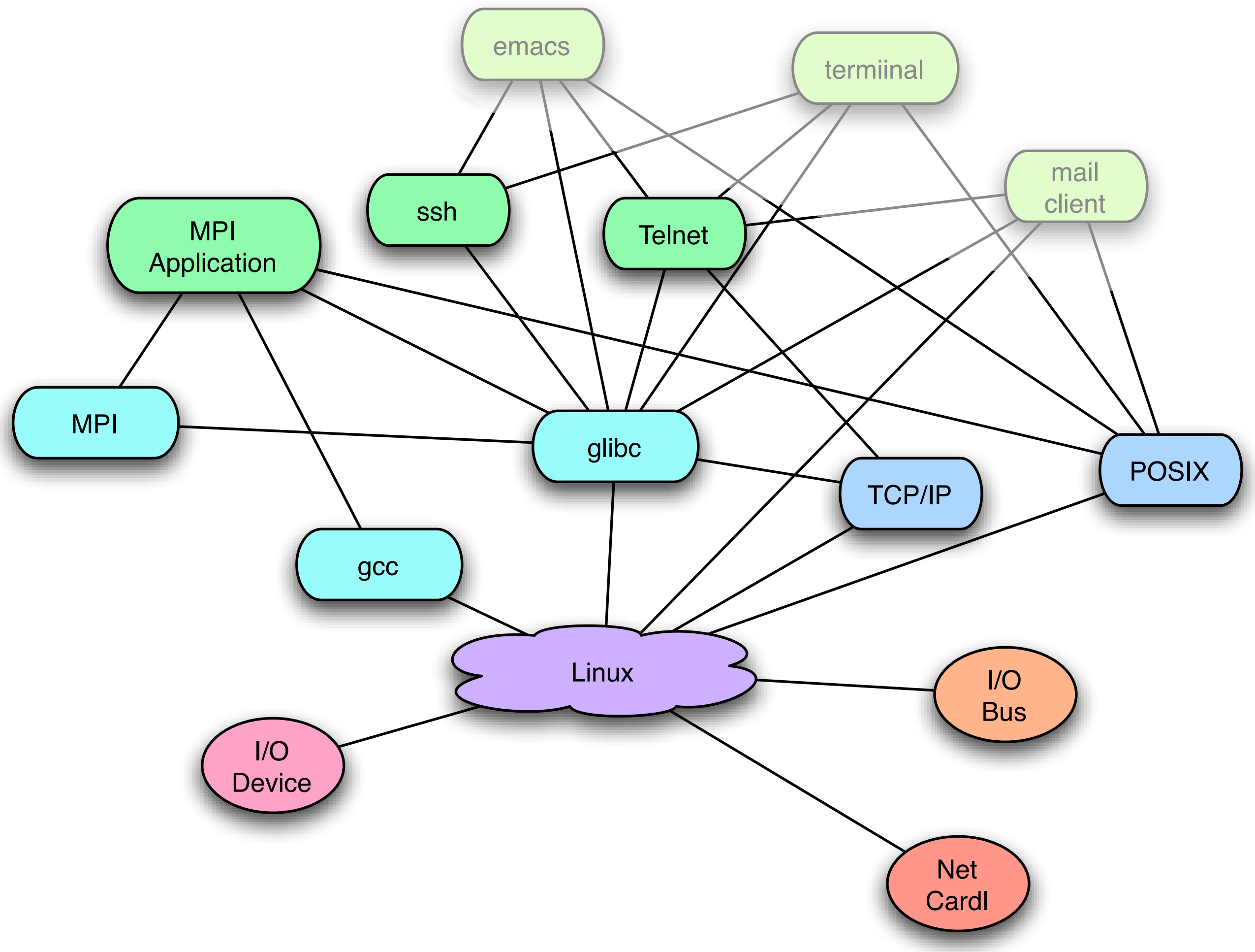


What does Linux do?

- Provides a wide range of services
  - libraries
  - development environment
  - work environment
- Works on a wide range of hardware
  - graphics cards
  - I/O buses
  - flaky stuff.....
- Hourglass design



What does Linux do in HPC?



- Don’t really have that many devices
  - No disks
  - none of it is flaky :)
- Must be the services
  - Probably not mail, emacs, or the terminal emulator....
  - “Real men read their mail on a Paragon”

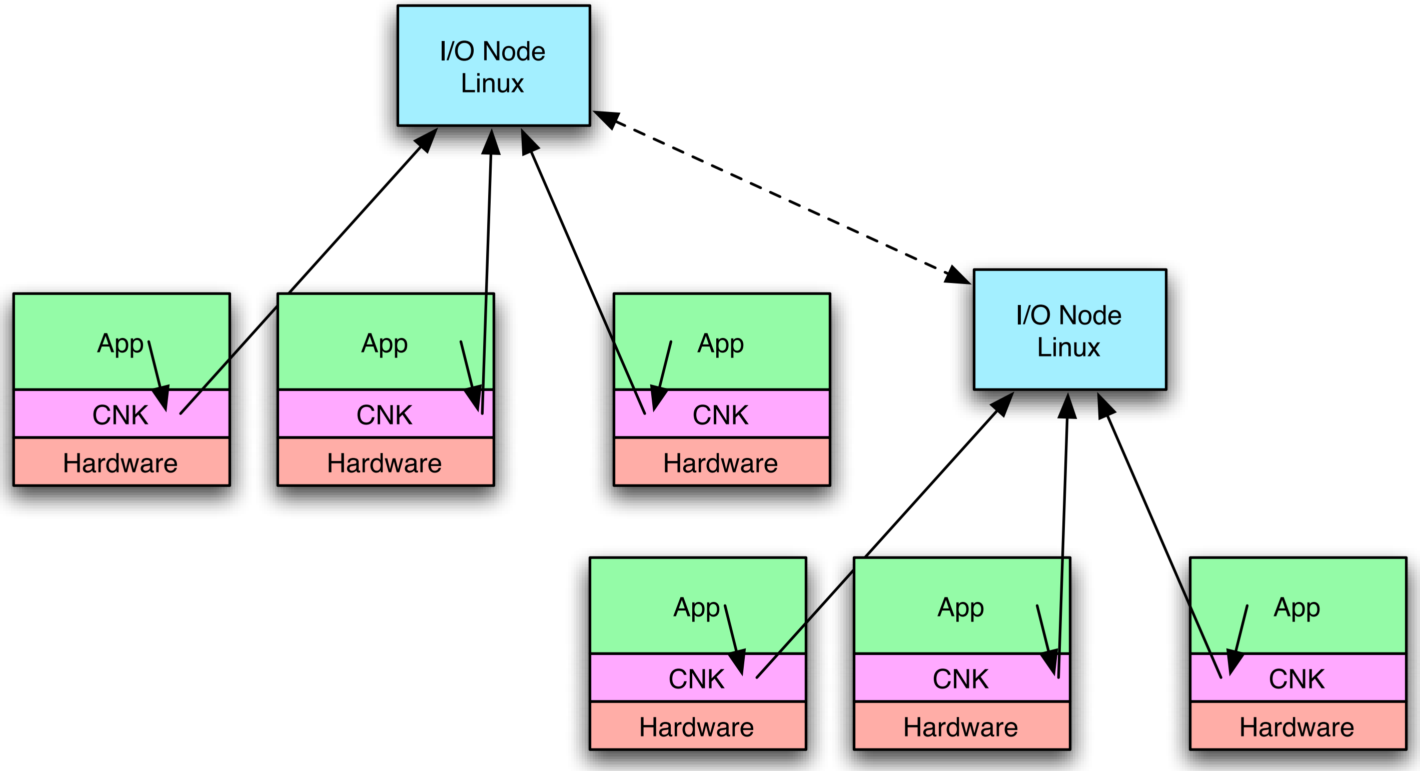
Lightweight Linux?

I’m busy planning to rule the world!

Well, good luck with that.

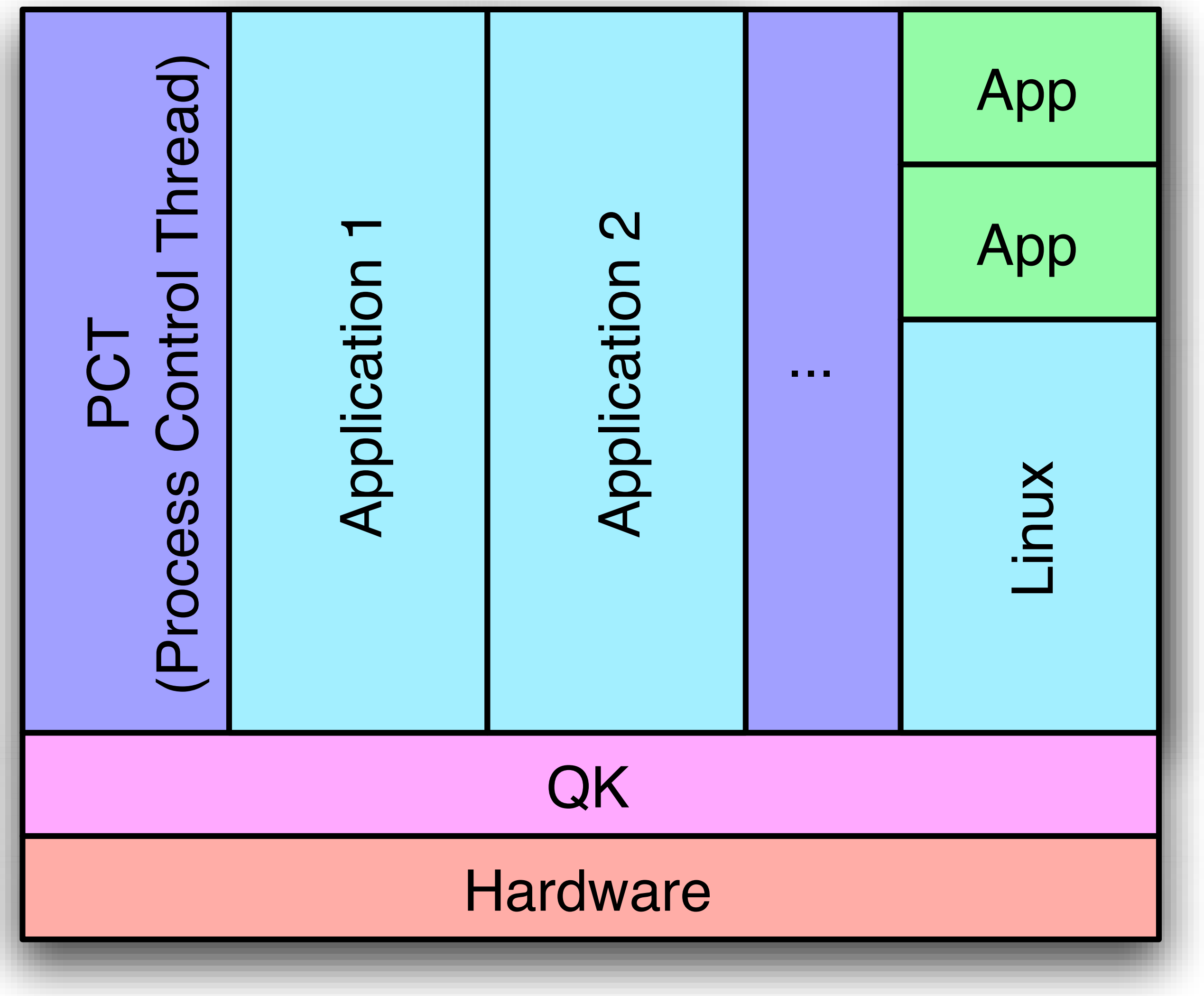
Running Linux on BG/L

- Seems like a "no-brainer"
  - some people will tell you that BG/L already runs Linux....
- It's not.....
  - "exec" is reasonable, but what does "fork" mean?
  - what is the right tradeoff for resources allocated to Linux?
  - Is that really Linux on the I/O nodes?



Linux on Catamount

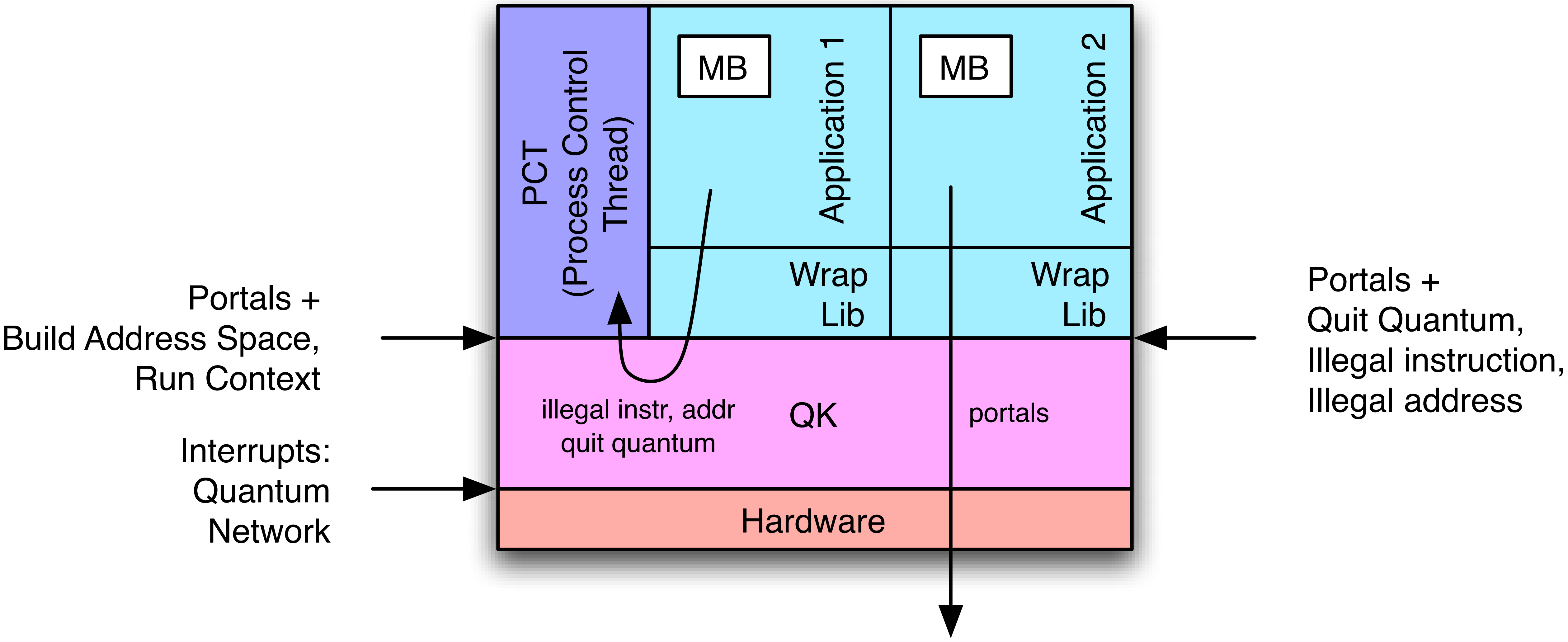
- Basic idea
  - QK == Xen
  - PCT == Dom 0
- QK virtualization
  - PCT builds address spaces
  - PCT can run contexts
  - Portals for network
- Use XenoLinux
  - emulate Xen hypercalls
  - no mod of XenoLinux



Xen Hypercalls

Hypercall	Meaning
set_callbacks	normal and “failsafe” handlers
sched_op_new	yield, block, shutdown, poll
mmu_update	update page table entries
stack_switch	change the stack
fpu_taskswitch	next use of FPU faults
memory_op	increase/decrease memory allocation
event_channel_op	inter-domain event-channel mgmt
physdev_op	BIOS Replacement

Catamount Mechanisms



PCT
(Process Control
Thread)

MB

Application 1

MB

Application 2

Wrap
Lib

Wrap
Lib

Portals +
Build Address Space,
Run Context

Portals +
Quit Quantum,
Illegal instruction,
Illegal address

illegal instr, addr
quit quantum

QK

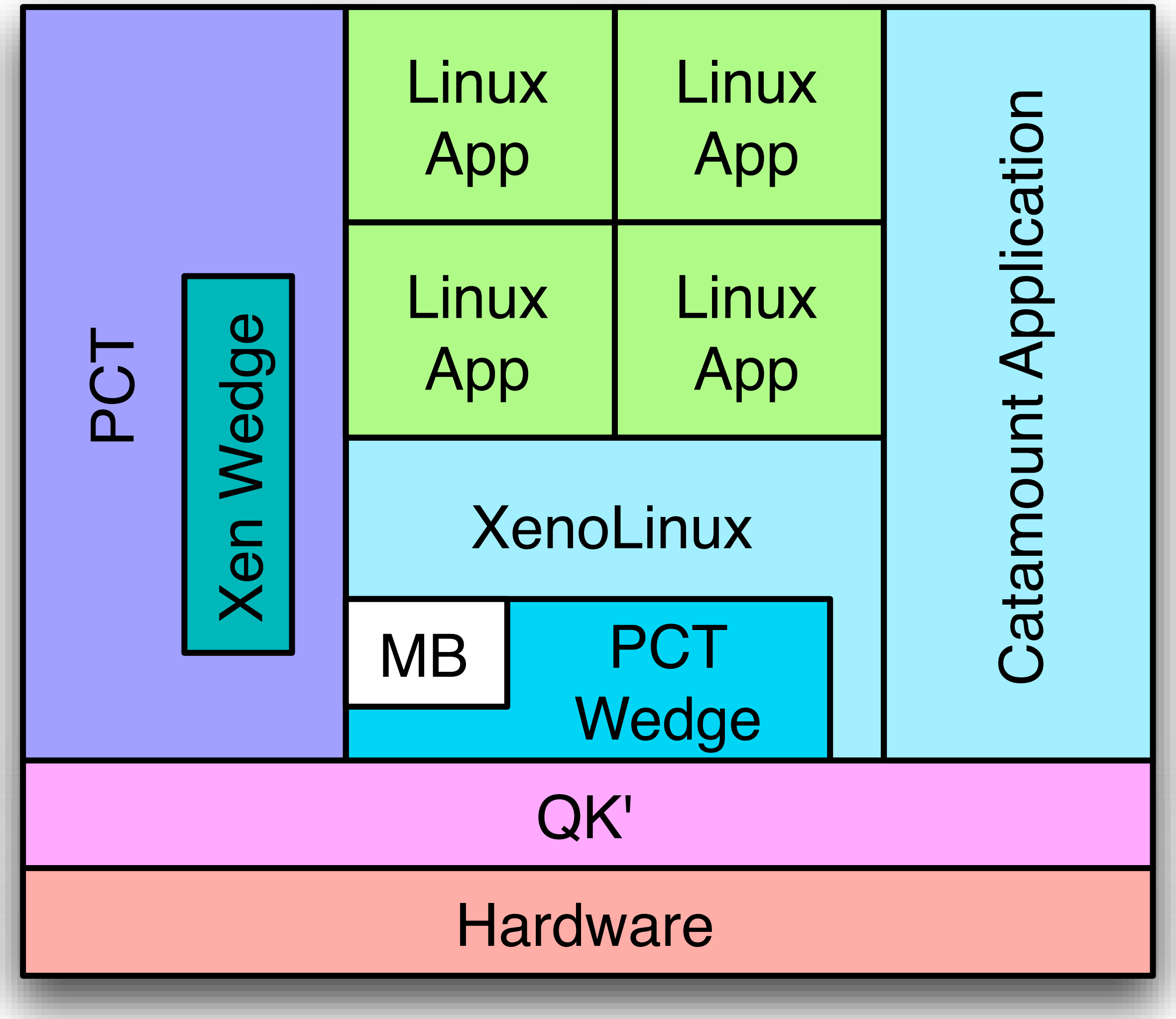
portals

Interrupts:
Quantum
Network

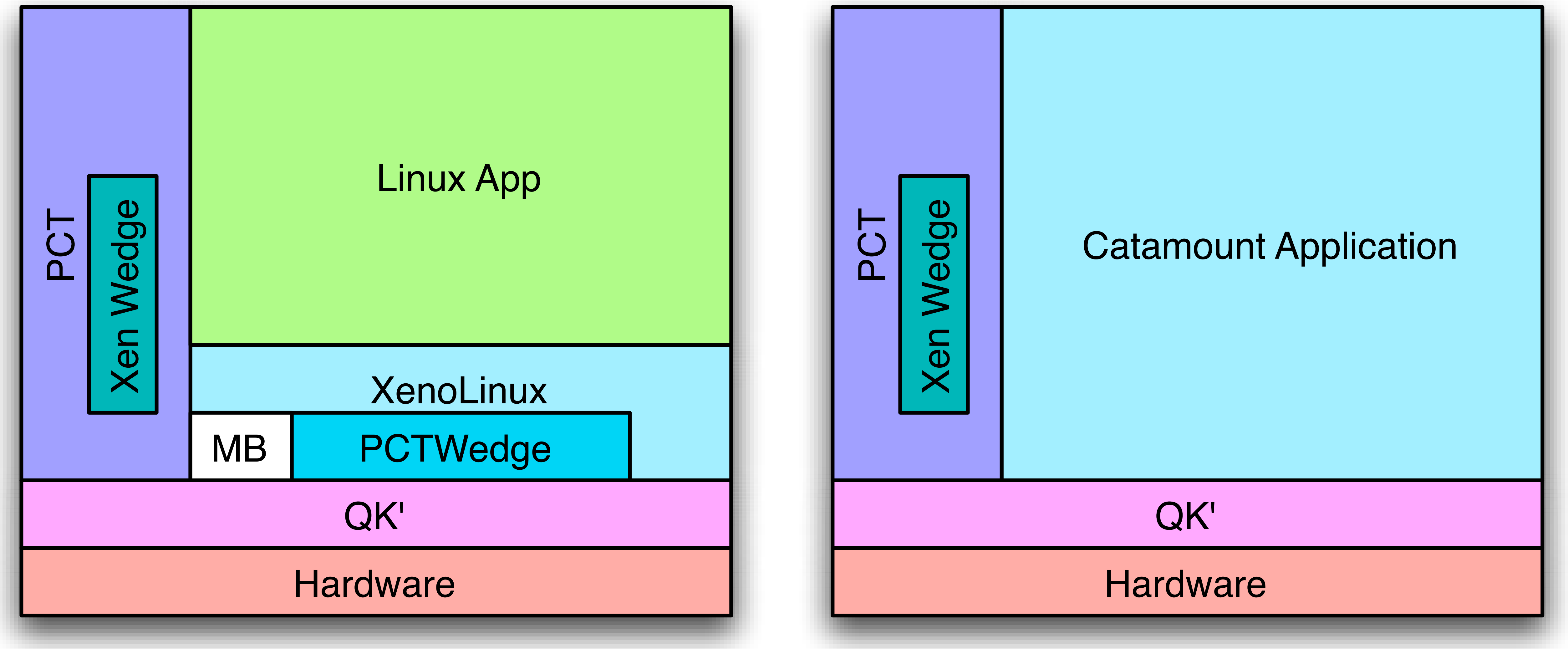
Hardware

A more realistic picture

- Start with XenoLinux
  - minimize modifications
  - build a wedge to provide QK interface
  - wedge could support page table construction
- Extend PCT and QK to support XenoLinux
  - minimize impact on Catamount applications
  - minimize changes to QK



Space Sharing



Never forget that the real goal is to run a single application per node (multiple processes, multiple threads)

Why Linux on Catamount?

- Linux is **not** free
  - Initial port and optimization
  - Linux evolves and requires updates
  - Does “lightweight” Linux exist?
- Catamount currently works and scales
  - not clear that Linux will scale
  - Catamount doesn’t evolve :) :)
- Use XenoLinux on Catamount
  - XenoLinux will evolve: evolve wedge, then PCT; QK only when necessary
  - Minimal number of supported code bases

FAST-OS

Forum to Address Scalable Technology for runtime and Operating Systems

Projects

	Activity
Colony	Virtualization on minimal Linux with SSI services
Config	Combine micro services to build app specific OS
DAiSES	Adaptation of OS based on Kperfmon & Kerninst
K42	Enhance applicability of K42 for HEC OS research
MOLAR	Modules to config and adapt Linux + RAS & fSM
Peta-Scale SSI	Intersection of big (SMP) and small (node) kernels
Right-Weight	Build application specific Linux/Plan 9 kernels
Scalable FT	Implicit, explicit, incremental checkpointing & resilience
SmartApps	Vertical integration between SmartApps and K42
ZeptoOS	Ultralight Linux, collective runtime, measure & FT

FAST-OS

	Virtualization	Adaptability	Usage Models	Metrics	Fault Handling	Common API	SSI	Collective RT	I/O	OS Noise	Linux
Colony	H		M		H	M	H		M	H	✔
ConfigOS	H	M	H					M	M	M	
DAiSES		H		H		M					✔
K42		H		H		H	M			M	
MOLAR		H	H	H	H			M		M	✔
Peta-Scale SSI			H		H		H		H	H	✔
Rightweight		M		H			M		M	H	½✔
Scalable FT					H			M	H	M	✔
SmartApps	M	H		H		M					
ZeptoOS			H	H	H			H		H	✔

H	High
M	Medium

Partners

	Lead	Academic	Industrial
Colony	LLNL	UIUC	IBM
Config	SNL	UNM, Caltech	
DAiSES	UTEP	Wisconsin	IBM
K42	LBNL	Toronto, UNM	IBM
MOLAR	ORNL	LaTech, OSU, NCSU	Cray
SSI	ORNL	Rice	HP, CFS, SGI, Intel
Right-Weight	LANL		Bell Labs
Scalable FT	PNNL	LANL, UIUC	Quadrics, Intel
SmartApps	Texas A&M	LLNL	IBM
ZeptoOS	ANL	Oregon	

FAST-OS

- PI meeting/workshop (open meeting)
  - with USENIX in Boston, May 30 & 31
- http://www.cs.unm.edu/~fastos
- Most recent issue of ACM OSR

> "Linux's cleverness is not in the software, but in the development model"
>
> Rob Pike, "Systems Software Research is Irrelevant," 2/2000